

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Recap and Checkpoint



Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Recap



So far we've learned about:

- Machine learning
- The ML pipeline
- Course projects
- Amazon SageMaker
- How to formulate a ML problem from a business problem

© 2022 Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



Navigate here to access Checkpoint #1:

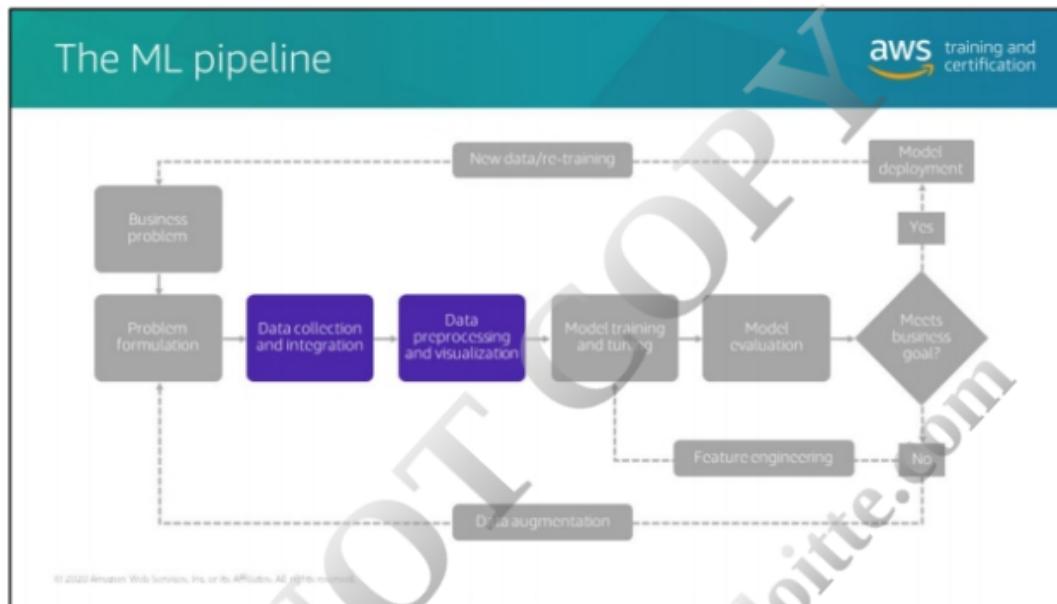
https://amazonmr.au1.qualtrics.com/jfe/form/SV_7aH7TW8Dwndo9gN

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



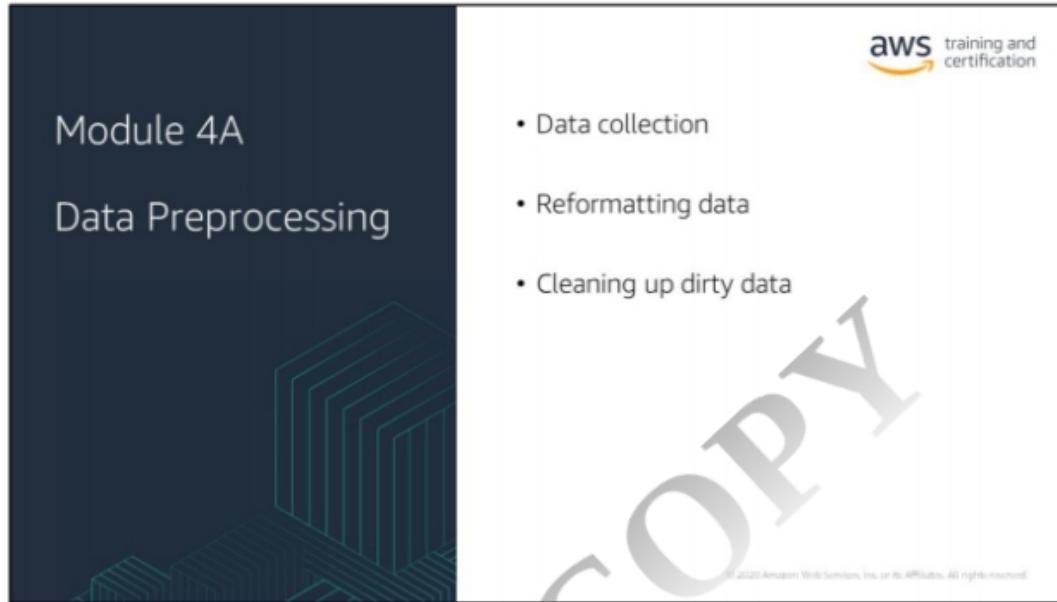
DO NOT COPY
amipandit@deloitte.com

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



Now that we've explored how to turn business problems into ML problems and have a better understanding of how to define success metrics, examine and choose your data, and choose your algorithm – it's time to start getting that data ready. This module covers how to preprocess your data and get it ready for training your model.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



The slide has a dark blue header section containing the title 'Module 4A' and 'Data Preprocessing'. Below this is a decorative graphic of green 3D bars. The main content area is white with the AWS training and certification logo at the top right. It contains a bulleted list of three items: 'Data collection', 'Reformatting data', and 'Cleaning up dirty data'. A large, diagonal watermark reading 'DO NOT COPY' and 'amipandit@deloitte.com' is overlaid across the slide.

- Data collection
- Reformatting data
- Cleaning up dirty data

©2022 Amazon Web Services, Inc. or its Affiliates. All rights reserved.

This module will focus on the next stage of the ML pipeline: Data preprocessing. This module covers data collection, reformatting data and techniques for cleaning up dirty data.

Note here that data visualization is woven throughout this module as, in reality, visualization and various preprocessing techniques happen simultaneously throughout this phase of the pipeline.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

You're going to have data from many different sources



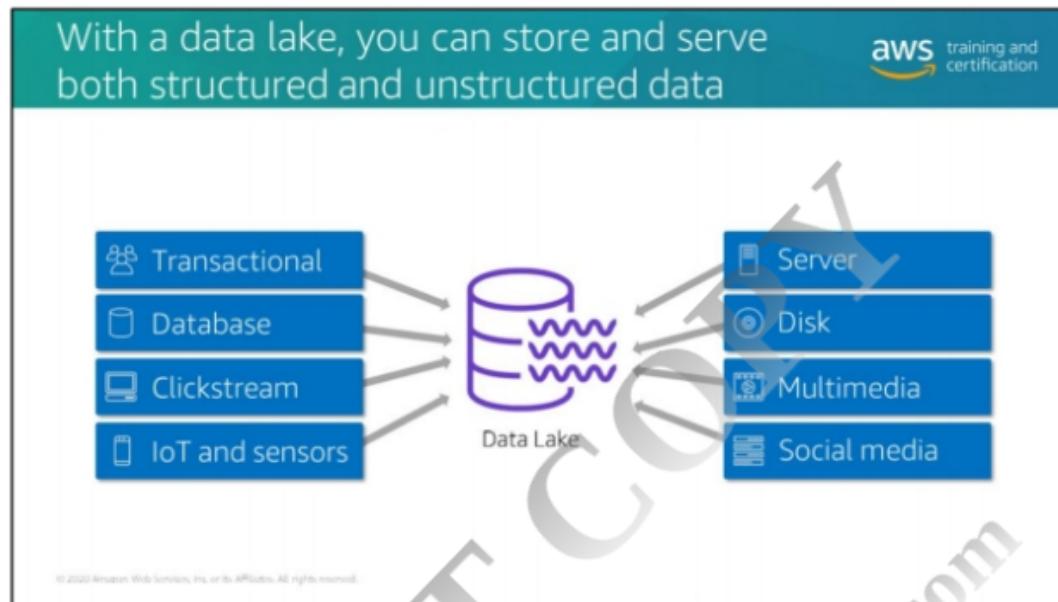
© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

 Transactional	 Server
 Database	 Disk
 Clickstream	 Multimedia
 IoT and sensors	 Social media

So we know what kind of data we need and how to do things like generate **labels** for datasets that aren't already labeled. Next, the question becomes: do we have access to this ideal data, what form is it in, and where is it currently stored? In real life, **ML** projects have data coming in from multiple sources and representing different types of data, including structured, semi-structured, and unstructured data. Data can come from a CSV file or a traditional database, for instance. When using the AWS Cloud, it could come from AWS storage services like Amazon Redshift, Amazon DynamoDB, Amazon RDS, or Amazon Glacier.

Organizations are often challenged to access and analyze their data when it's stored in different formats across various, disparate locations. With data continuously being collected from a variety of sources, if not properly addressed, this challenge only grows bigger as organizations age and grow. A lack of simplified access to data creates workflow bottlenecks as employees often need to request help from IT to access the information required to build, train, and deploy machine learning models. To address these challenges, organizations are forced to seek out a solution that offers a single source of truth, that is readily accessible when, and to the employees that need it. This is often difficult and costly to maintain as an on-premises solution.

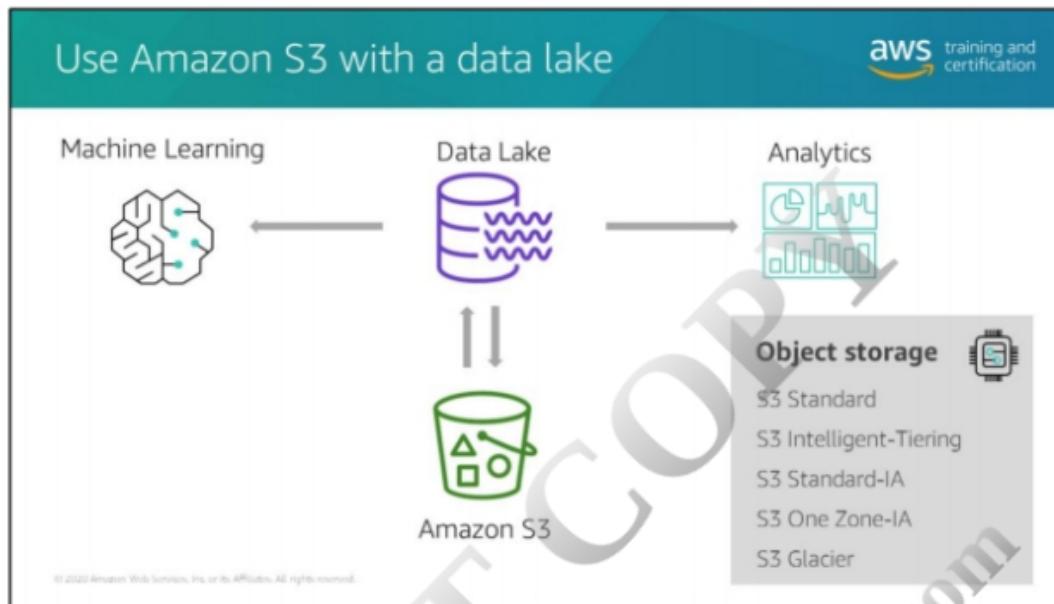
Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



A data lake architecture can provide a solid foundation on which to build a solution to this challenge. A data lake allows you to store massive amounts of data in a central repository so it's readily available to be categorized, processed, enriched, and consumed by diverse groups within an organization.

There's no silver bullet to building a data lake, however. In most cases, building a data lake entails the use of dozens of technologies, tools, and environments, including data from third parties. When done right, a data lake can open the door to a whole new set of advanced analytics, facilitating data science and machine learning.

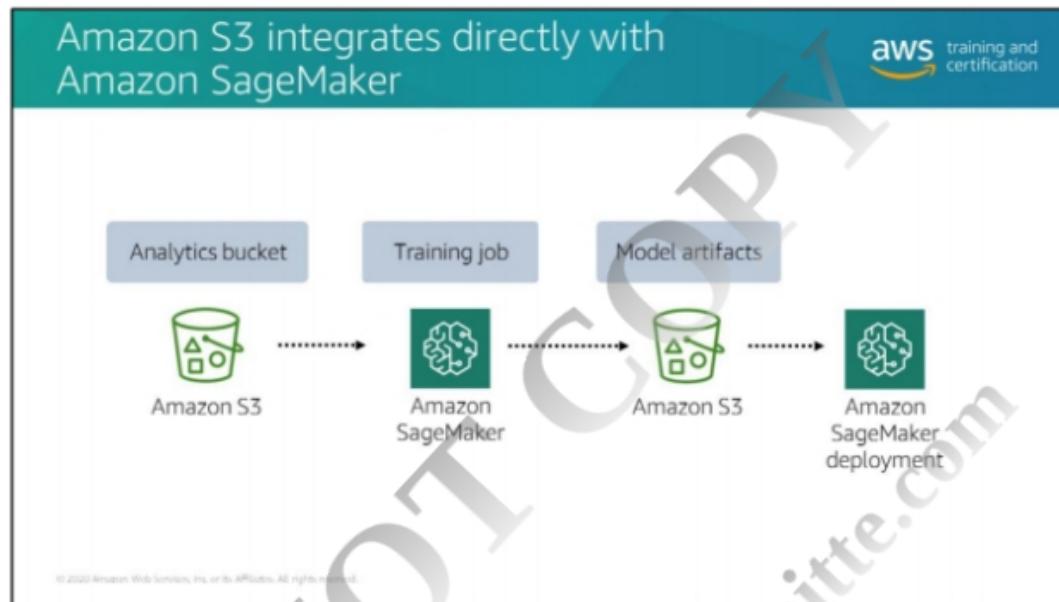
Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



The data lake foundation is designed to provide you with a single source of truth, with additional features including data submission, ingest processing, aggregation, analysis, and searching capabilities.

Amazon S3 is the preferred storage option with a data lake for data science processing on AWS. Amazon S3 provides highly durable storage and seamless integration with various data processing services and ML platforms on AWS. It can be used as "one source of truth" storage for most AWS ML services.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



So why should you use Amazon S3 to address your machine learning needs?

First, you can use Amazon S3 to store your training datasets that you will use for your machine learning models. These files can be delivered directly to your Amazon SageMaker training jobs from Amazon S3.

Second, Amazon S3 can also be used for storing objects like trained ML models or images or videos that need preprocessing before predictions including workloads with large amounts of user generated content, such as video or photo sharing.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



Alternatively, if your training data is already in Amazon Elastic File System (Amazon EFS), we recommend using that as your training data source. Amazon EFS has the benefit of directly launching your training jobs from the service without the need for data movement, resulting in faster training start times. This is often the case in environments where data scientists have home directories in Amazon EFS and are quickly iterating on their models by bringing in new data, sharing data with colleagues, and experimenting with including different fields or labels in their dataset.

For example, a data scientist can use a Jupyter notebook to do initial cleansing on a training set, launch a training job from Amazon SageMaker, then use their notebook to drop a column and re-launch the training job, comparing the resulting models to see which works better.

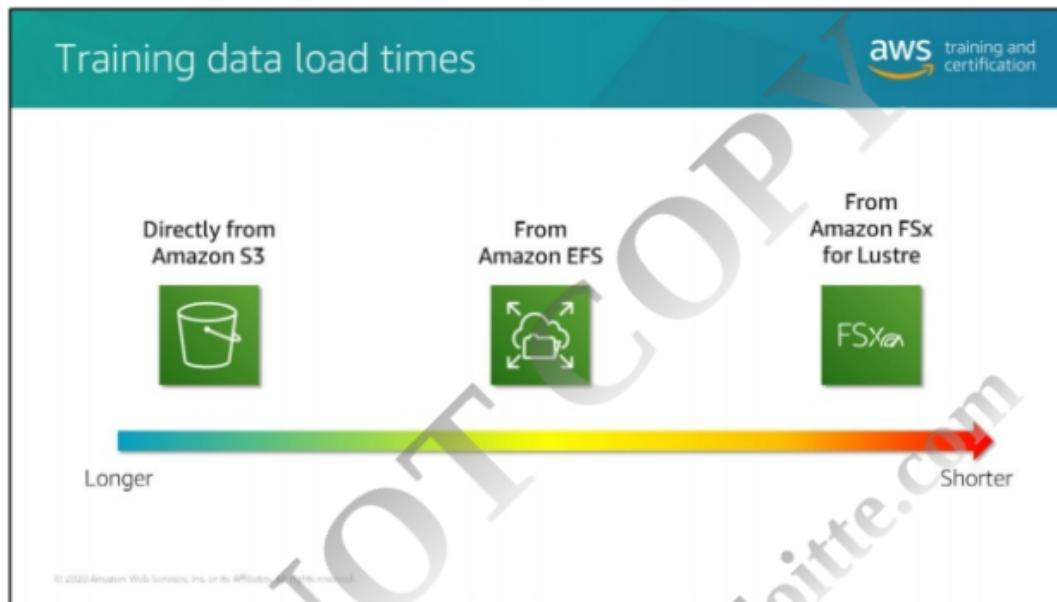
Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



When your training data is already in Amazon S3 and you plan to run training jobs several times using different algorithms and parameters, consider using Amazon FSx for Lustre, a file system service. FSx for Lustre speeds up your training jobs by serving your Amazon S3 data to Amazon SageMaker at high speeds. The first time you run a training job, FSx for Lustre automatically copies data from Amazon S3 and makes it available to Amazon SageMaker. You can use the same Amazon FSx file system for subsequent iterations of training jobs, preventing repeated downloads of common Amazon S3 objects.

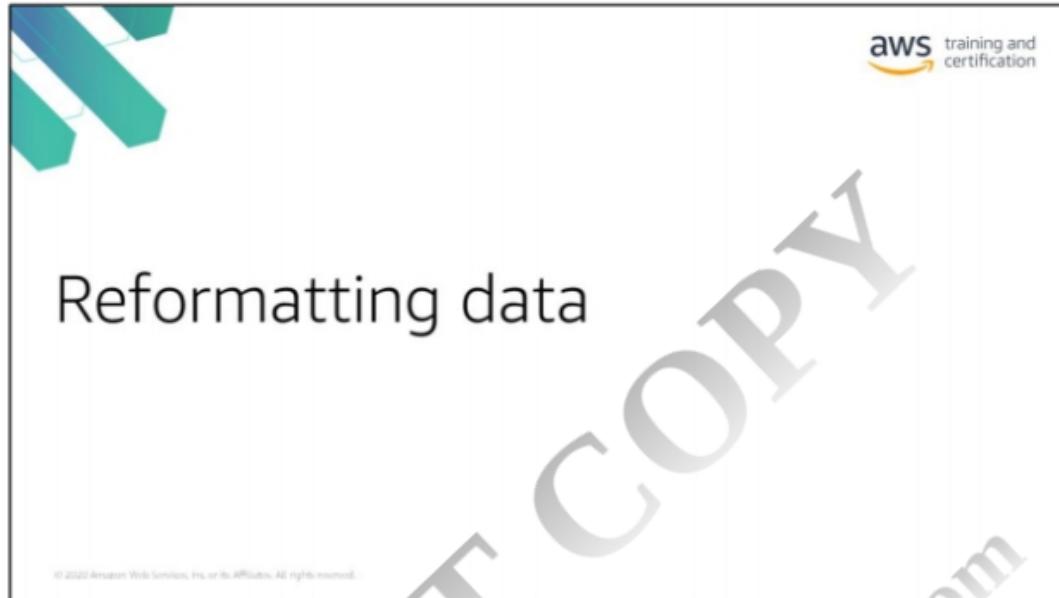
You can shut down Lustre whenever your workloads complete, and only use it during training.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



When it comes to running ML workloads, you want to be cognizant of the training load time. Consider loading images, for example, if you were to do a quick comparison of Amazon S3 and Amazon EFS, you'd notice that Amazon S3 is the cheaper option for data storage. Storing the image datasets which are large in size in to Amazon S3 makes sense when you don't have a training job time constraints that you need to take into consideration. When you have training job time constraints, you want your data to be as close to instances as possible while providing the high throughput and sub-second latencies. You can use Amazon FSX for Lustre and Amazon EFS for these types of workloads.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Example: Raw fraudulent transaction data 

Example of raw data

```
"Customer,DateOfTransaction,Vendor,ChargeAmount,WasThisFraud"  
"ABC, 10/5, Store1, 10.99, No...."
```

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

Think back to our fraudulent transaction scenario and take a look at the example on the screen. Before you can start running statistics on your data to better understand what you're working with, you have to ensure it's in the right format to be analyzed. So how do you get your data into the proper format? You should begin by asking yourself questions like these:

- How would I fix the problem with this data?
- What tools should I use to help me fix it?

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

The slide has a dark blue header with the title "Using Pandas: Dataframes". Below the header is a decorative graphic of green 3D bars. The main content area is white with a light gray watermark reading "amipandit@deloitte.com" diagonally across it. At the top right is the AWS logo with the text "training and certification". A blue speech bubble contains the JSON string: "Customer:ABC, DateOfTransaction:10/5 ,Vendor...". Below the bubble is the heading "Formatted data". A table follows:

Customer	Data of transaction	Vendor	Charge amount	Was this fraud?
ABC	10/5	Store 1	10.99	No
DEF	10/5	Store 2	99.99	Ycs
GHI	10/5	Store 2	15.00	No
JKL	10/6	Store 2	99.99	?
MNO	10/6	Store 1	99.99	Ycs

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

Pandas, an open source Python library, can be used for this particular data reformatting. Pandas reformats data from various formats like CSV, JSON, Excel, Pickle, and others into a tabular representation, presenting it in rows and columns.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

The diagram shows the construction of a DataFrame from two Series. On the left, there is a dark blue background with the text "Using Pandas: Dataframes" and a stylized 3D bar chart graphic. To the right, the AWS training and certification logo is at the top. Below it, two Series are shown as tables:

Series	
	Cats
0	5
1	77
2	5
3	3
4	0

Series	
	Dogs
0	20
1	1
2	6
3	6
4	13

A large plus sign (+) is positioned between the two Series tables, followed by an equals sign (=). To the right of the equals sign is a final DataFrame table:

DataFrame		
	Cats	Dogs
0	5	20
1	77	1
2	5	6
3	3	6
4	0	13

© 2020 Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The tabular format that Pandas puts your data into is called a *dataframe*. A dataframe is made up of many *series*, which are essentially columns capable of holding any data type. The axis labels are referred to as the *index*.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Using Pandas:
Dataframes

aws training and certification

You can:

- Calculate statistics
- Clean your data
- Visualize it
- Store the cleaned and transformed data back into its original format (e.g., a CSV file)

© 2022 Amazon Web Services, Inc. or its Affiliates. All rights reserved.

With your data in a Pandas dataframe, you can calculate statistics, clean your data, visualize it, and even store the cleaned and transformed data back into its original format (e.g., a CSV file).

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Import Pandas and load your dataset into a dataframe

```
aws training and certification
In [1]: import pandas as pd
In [2]: df = pd.read_csv('YourDataSet.csv')

You can import from other file formats, such as:
• JSON
• Parquet
• HTML
• SQL
• and more (link in the notes)
```

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

To get started, you can import Pandas using the code above. Then, import the dataset from a local CSV file, for instance, and then show the first few rows of the dataframe. You can also import from other file formats. The complete list of compatible file formats is here: https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

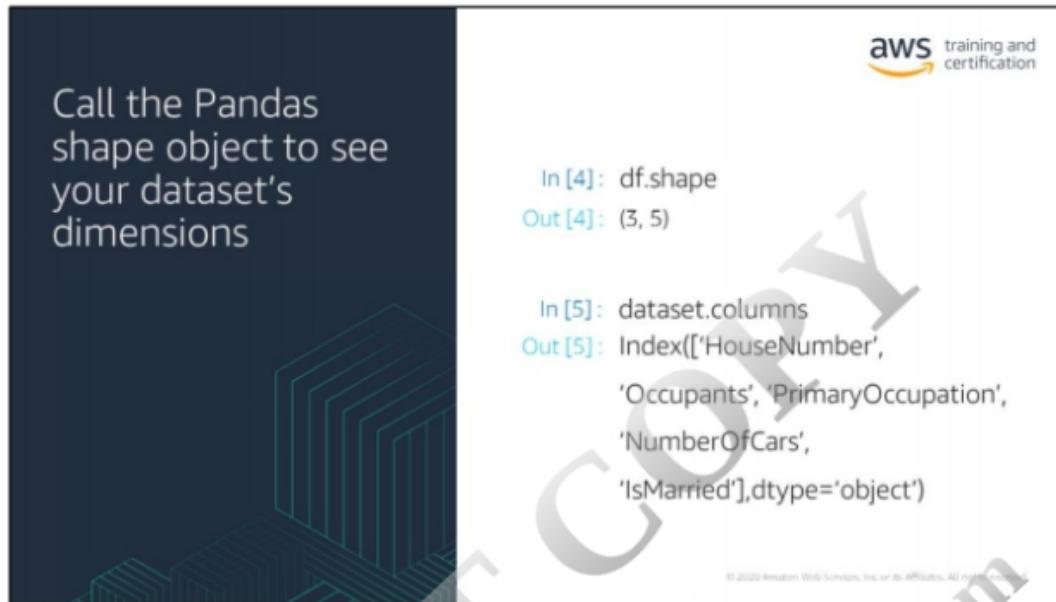
The slide features a dark blue background with a subtle circuit board pattern at the bottom. On the left, white text reads: "Make sure your dataframe is constructed properly by importing the first few rows". On the right, there is a screenshot of a Jupyter Notebook cell. The cell starts with "In [1]: import pandas as pd", followed by "In [2]: df = pd.read_csv('YourDataSet.csv')". Then, "In [3]: df.head(3)" is run, and the output "out [3]:" shows a Pandas DataFrame:

Customer	Date of transaction	Vendor	Charge amount	Was this fraud?
ABC	10/5	Store 1	10.99	No
DEF	10/5	Store 2	99.99	Yes
GHI	10/5	Store 2	15.00	No

At the bottom of the slide, a small watermark reads "© 2020 Amazon Web Services, Inc. or its Affiliates. All rights reserved."

Then, import the dataset from a local CSV file, for instance, and then show the first few rows of the dataframe.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



With Pandas, for instance, you can call the shape object to see your dataset's dimensions. Shape returns the number of rows and columns. Columns returns the names and types of your columns, and Rows does the same for rows.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

The slide has a teal header bar with the title "More libraries for reformatting your data" and the AWS training and certification logo. Below the header, there are four dark blue rectangular boxes containing the names of the libraries: "NumPy", "Scikit-Learn", "Matplotlib", and "Seaborn". Underneath each library name is a snippet of Python code:

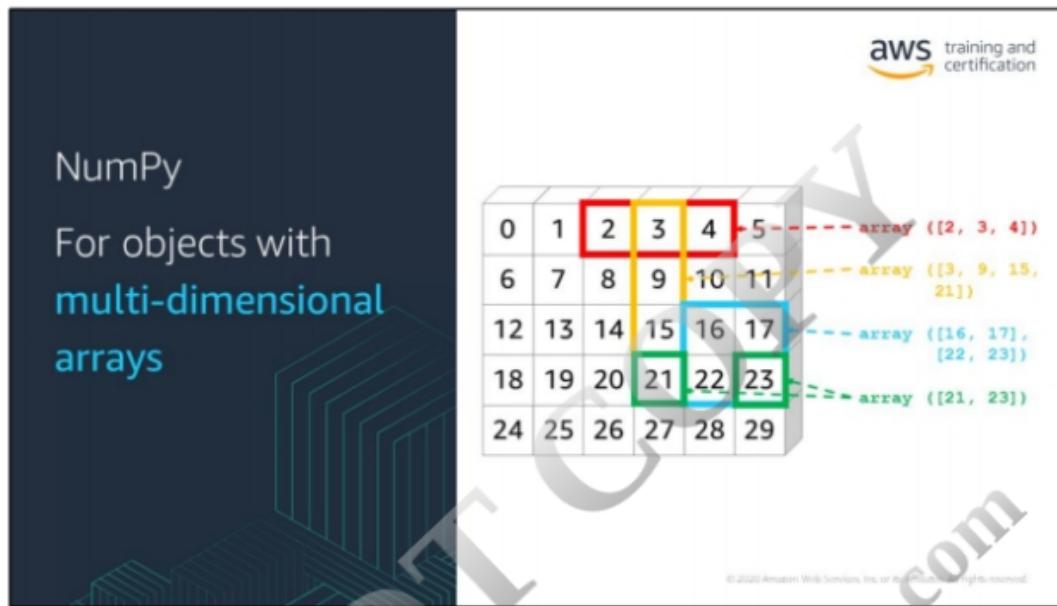
```
import numpy as np      from sklearn import datasets    import matplotlib.pyplot as plt    import seaborn as sns
```

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

In machine learning, libraries are tools containing a premade set of routines and functions that enable you to format different types of data, like numerical or image data.

Here's an overview of common tools used at this preprocessing phase of the ML pipeline.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

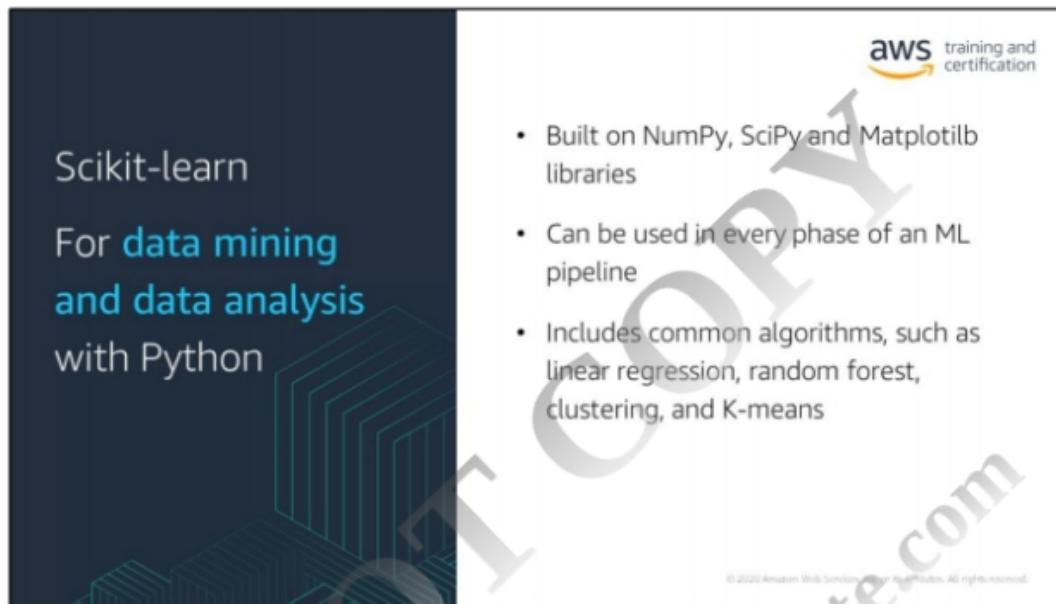


NumPy: NumPy is a Python library that serves as a fundamental package for scientific computing while working with Python. It's a general-purpose array-processing package that provides high-performance multidimensional array object and the tools for working with these arrays. Its features include:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, and random number capabilities

Given its features, NumPy is a common tool used at the preprocessing phase of the ML pipeline as it can help you reformat your data and run a variety of statistics on that data to better understand it.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



Scikit-learn: Scikit-learn is an open source Python library that includes various tools for data mining and data analysis. It is built on NumPy, SciPy and Matplotlib libraries. It can be used in every phase of a machine learning pipeline including data preprocessing. We'll circle back to Scikit-learn during the model training phase of the pipeline as well, as the library includes supervised algorithms like linear regression and random forests as well as unsupervised algorithms like clustering and K-means.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Matplotlib

A **visualization library** for Python used for two-dimensional plots of NumPy arrays

- Can visualize your data in a number of ways, including line and bar charts, scatterplots, and histograms
- Works directly from Python scripts, Jupyter notebooks, IPython shells, and other platforms

aws training and certification

© 2020 Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Matplotlib: Matplotlib is a visualization library in Python used for two-dimensional plots of NumPy arrays. With Matplotlib you can visualize your data in a number of ways, including line and bar charts, scatterplots, and histograms. You can leverage the library directly in your Python scripts, Jupyter notebooks, IPython shells, and other platforms.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Seaborn
Another
visualization
library for Python

- Built on top of Matplotlib
- Closely integrated with Pandas DataFrames
- Provides a high-level interface for drawing attractive and informative statistical graphics

© 2022 Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Seaborn: Seaborn is another Python data visualization library. It's built on top of Matplotlib and is closely integrated with Pandas DataFrames. It provides a high-level interface for drawing attractive and informative statistical graphics.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



Now that you have reformatted your data, it's time to begin cleaning it up.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Example data for training a model to predict future flu outbreaks							
Participant ID	Age	Income	Education	State	Flu shot	Outbreak zone?	
123456	39	45,000/year	4 yr degree	New York	Yes	No	
123457	23	1,500/month	Baccalauréat	Minnesota	No	Yes	
123458	78		Masters/PhD		Yes	Yes	
123459	20	3,000,000/year	HS diploma	California	No	No	
123460	154	53,000/year		Masters/PhD			
***	***	***	***	***	***	***	

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

For the purposes of demonstrating how we're going to pre-process some data, here is an example of a few fictional rows of data we've created for the purposes of this course. This dataset would be used to predict future flu outbreaks by training the model on historical data about people, including demographic data as well as whether they got a flu shot last year, along with some corresponding data about their location and whether it was subject to a flu outbreak during that year's flu season.

Now the problem with this dataset is pretty clear. We've got a problem with what's called *dirty data*.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Example data for training a model to predict future flu outbreaks						
Participant ID	Age	Income	Education	State	Flu shot	Outbreak zone?
123456	39	45,000/year	4 yr degree	New York	Yes	No
123457	23	1,500/month	Baccalauréat	Minnesota	No	Yes
123458	78		Masters/PhD		Yes	Yes
123459	20	3,000,000/year	HS diploma	California	No	No
123460	154	53,000/year		Masters/PhD		
***	***	***	***	***	***	***

Languages,
grammar, or
slang that's
not expected

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data can be messy in several ways. For instance, maybe your algorithm expects to see data written in English, but there are some words in your data set from different languages. Or maybe there are special characters in some of the words, or even just a lot of space in between words. The key is to make sure you are standardizing your data. If your algorithm requires English, make sure it's all in English.

The same goes for grammatical structure. For example, convert your text data into all lowercase so the same word isn't treated as two different words just because of its capitalization.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Example data for training a model to predict future flu outbreaks							
Participant ID	Age	Income	Education	State	Flu shot	Outbreak zone?	
123456	39	45,000/year	4 yr degree	New York	Yes	No	Different scales in the same column
123457	23	1,500/month	Baccalauréat	Minnesota	No	Yes	
123458	78		Masters/PhD		Yes	Yes	
123459	20	3,000,000/year	HS diploma	California	No	No	Common examples: currencies, measurement, and time
123460	154	53,000/year		Masters/PhD			
***	***	***	***	***	***	***	

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

Your data set might also include data that is on very different scales. For example, here we have one column called *Length*, but that column has different units for data, like kilometers, meters, and miles. This is a common occurrence in many numerical data sets, especially if your dataset is a result of merging data from multiple sources.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Example data for training a model to predict future flu outbreaks

aws training and certification

Participant ID	Age	Income	Education	State	Flu shot	Outbreak zone?
123456	39	45,000/year	4 yr degree	New York	Yes	No
123457	23	1,500/month	Baccalauréat	Minnesota	No	Yes
123458	78		Masters/PhD		Yes	Yes
123459	20	3,000,000/year	HS diploma	California	No	No
123460	154	53,000/year		Masters/PhD		
***	***	***	***	***	***	***

You could write simple code to find these errors and replace them

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

For both of these kinds of examples, once you identify common errors like these, you can write some simple code that finds these errors and changes them. For instance, your code could look for anything in the income column that says "/month" or "/per month" and multiply the value by 12. Or for unexpected languages, it could replace the entries with the word in the expected language for your data.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Example data for training a model to predict future flu outbreaks

aws training and certification

Participant ID	Age	Income	Education	State	Flu shot	Outbreak zone?
123456	39	45,000/year	4 yr degree	New York	Yes	No
123457	23	18,000/year	HS diploma	Minnesota	No	Yes
123458	78		Masters/PhD		Yes	Yes
123459	20	3,000,000/year	HS diploma	California	No	No
123460	154	53,000/year		Masters/PhD		
***	***	***	***	***	***	***

More than one feature in the same column

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

An even messier example is when you have a column of data that has multiple features represented. For instance, the stat column also has data from other columns in it. In this situation, you have to reshape the data so that each specific feature will be in its own column.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Example data for training a model to predict future flu outbreaks

aws training and certification

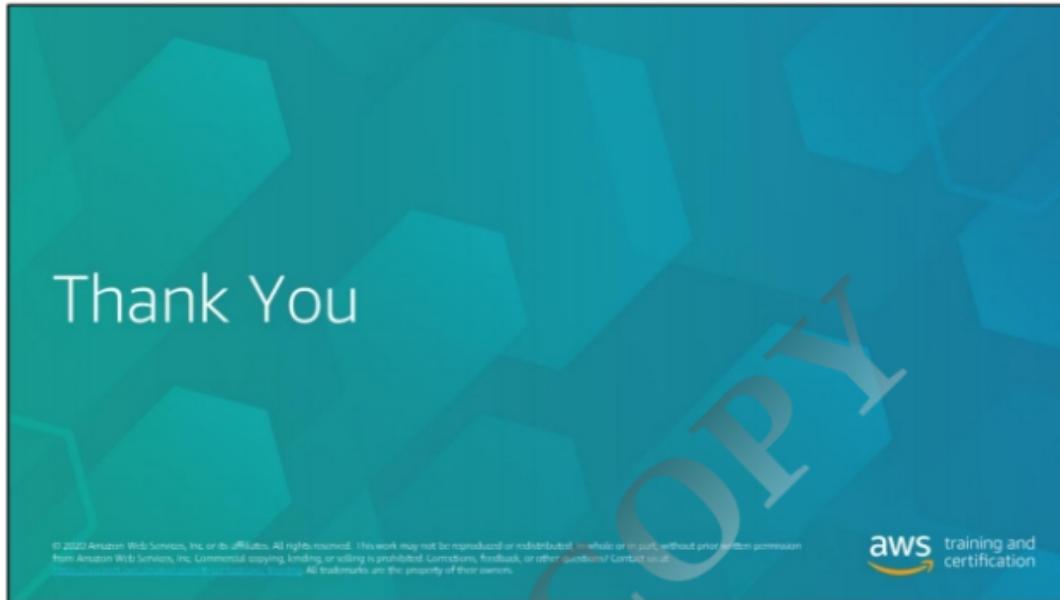
Participant ID	Age	Income	Education	State	Flu shot	Outbreak zone?
123456	39	45,000/year	4 yr degree	New York	Yes	No
123457	23	18,000/year	HS diploma	Minnesota	No	Yes
123458	78		Masters/PhD		Yes	Yes
123459	20	3,000,000/year	HS diploma	California	No	No
123460	154	53,000/year		Masters/PhD		
***	***	***	***		***	***

Get the data scientist to normalize this column

© 2022 Amazon Web Services, Inc. or its affiliates. All rights reserved.

To reshape the data so that the features are in their own column, you need the column to be normalized. If you're not familiar with this process that's ok, your data scientist can help.

Printed by: amipandit@deloitte.com. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



DO NOT COPY
amipandit@deloitte.com